Protistology 5 (2/3), 131–141 (2007/8)

Protistology

Macroarray-based gene expression analysis in *Tetrahymena pyriformis*

Bernhard F. Benkel , Scott Richmond, Jon Davoren, Michael Ivan, Ronald M. Teather and Robert J. Forster

Agriculture and Agri-Food Canada, Lethbridge Research Centre, Lethbridge, Alberta, Canada

Summary

A cDNA library was constructed for the aquatic ciliate *Tetrahymena pyriformis*, and one hundred and fifty-three clones were used for gene expression analysis in a filter array hybridization assay. Expression levels for individual genes ranged from easily detectable to below the detection threshold of the assay. Ten cDNAs showed relatively intense hybridization signals, and two of these were chosen for further analysis. One of these genes, Tp177, was characterized in detail, i.e. its transcribed and putative promoter regions were completely defined and analyses carried out to identify potential gene regulatory elements. Partial nucleotide sequences were derived from fifty-seven cDNA clones and compared to the GenBank sequence database. Twenty-six of these clones showed significant similarity to genes in GenBank (eleven of these represented ribosomal proteins), while the remainder, including clone Tp177, represented novel sequences.

Key Words: EST, cDNA array, transcript abundance, gene cloning, long range-inverse PCR, DNA sequencing, protozoa.

Introduction

The unicellular eukaryote, *Tetrahymena*, has a lengthy history as a valuable experimental model organism due to the ease with which it can be cultured and manipulated, and its amenability to genetic and molecular characterization. *Tetrahymena* has served as the primary model system for the discovery and investigation of a number of fundamental molecular and cellular phenomena including ribozymes and telomeres. In addition, *Tetrahymena* is a promising system for biotechnological applications. This is due in part to the ability of *Tetrahymena* cells to produce copious quantities of recombinant proteins (Shang et al., 2002). The development of transformation methodology for *Tetrahymena* (Gaertig and Gorovsky, 1992; Cassidy-Hanley et al., 1997) coupled with its capacity for high level, synchronous protein secretion has opened up the possibility of the use of this organism for industrial synthesis of recombinant proteins.

In order to harness the protein producing capacity of *Tetrahymena*, shuttle vectors will need to be constructed that allow for the introduction of recombinant gene constructs driving the expression of foreign genes. Among the essential components in the construction of such shuttle vectors are gene promoters that direct the high level transcription of

foreign protein-encoding sequences. The purpose of this study was to evaluate the expression levels of a set of *Tetrahymena pyriformis* genes during vegetative growth under normal culture conditions in order to identify genes with robust activity profiles. The promoters of such genes can be used directly in hybrid constructs to drive the expression of recombinant peptides. In addition, the characterization of essential gene regulatory elements within strong promoters will eventually facilitate the construction of synthetic gene expression cassettes for high level, regulated production of foreign proteins in *Tetrahymena*.

Materials and Methods

Culture and Harvest of Tetrahymena Cells

Commercially prepared cultures of *Tetrahymena pyriformis* were purchased from Boreal Laboratories Ltd (St. Catharines, Ontario, Canada). The cells were passaged weekly at a dilution of 1:100 in SSP medium (Orias et al., 2000) supplemented with 250 ug/ml each of penicillin G and streptomycin sulfate. Cells for nucleic acid isolation were harvested 4 days after passaging by centrifugation at 1,500 g for 5 min.

Genomic DNA Extraction

Genomic DNA was extracted from *T. pyriformis* cells using the Genomic-tip 20/G procedure (Qiagen, Mississauga, Ontario) according to the manufacturer's directions with the following modifications. The incubation period for purified nuclei in the G2 buffer/proteinase K solution was increased from 60 to 90 minutes. Longer incubation time aided in clearing the lysate and reduced blocking of the column. The DNA was precipitated from the eluate with isopropanol and pelleted at 3,838 g for 30 min at 4 °C. Following a final wash with 70% ethanol, the pellet was recentrifuged at 3,838 g for 20 min at 4 °C, dried, and resuspended in nuclease-free water.

RNA Extraction

Total RNA was prepared using Trizol reagent (GibcoBRL; currently Invitrogen Life Sciences, Burlington, Ontario, Canada) according to the manufacturer's protocol for RNA isolation from suspension-grown cells, with the following modifications. One mL of cell suspension was extracted with 4 mL of Trizol reagent and the phases separated by centrifugation at 3,838 g for 15 min at 4 ° C. The RNA was precipitated from the aqueous phase with isopropanol, pelleted by centrifugation at 3,838 g for 20 min at 4 °C, and washed with 75% ethanol. Total RNA was resuspended in nuclease-free water and polyadenylated RNA (poly-A+ RNA) was prepared using the PolyATract mRNA Isolation System (Promega Life Sciences, Madison, Wisconsin, USA) following the manufacturer's protocol for small-scale mRNA isolation.

cDNA Preparation and Cloning

Poly-A^+ RNA was used as the template for cDNA synthesis using the SMART kit (Clontech: currently BD Biosciences Clontech, Windsor, Ontario, Canada). Thirteen cycles of PCR amplification, carried out as specified in the SMART kit protocol, were used to generate double-stranded cDNA. An aliquot of the amplified cDNA was ligated to pGEM-T vector (Promega) and transformed into *E. coli* DH5α Max Efficiency competent cells (GibcoBRL) according to the manufacturer's directions.

RNA Abundance Analysis

Inserts from 153 randomly selected cDNA clones were amplified by PCR using the SMART primer (see Table 2), and 0.3-µg and 1.0-µg aliquots of each sample were prepared for blotting as described below. Only a single primer was required for PCR amplification of cDNAs since the SMART primer is incorporated at both ends of the insert during cDNA synthesis using the SMART kit. The 0.3 and 1.0 µg samples of PCR-amplified cDNA were brought to a final volume of 200 µL with TE (pH 8.0), mixed with twenty µL of 3 N NaOH, and incubated at 60 °C for 60 min. The samples were cooled to room temperature, 220 µL of 6x SSPE was added and mixed by vortexing, and the total samples were applied to Hybond-N+ membranes (Amersham Biosciences, Baie d'Urfé, Québec, Canada) using a Minifold II slot blot system (Schleicher & Schuell, Keene, New Hampshire, USA). DNA was fixed to membranes using a Stratalinker (UV crosslinker; Stratagene, La Jolla, California, USA).

Blots were hybridized with radiolabeled single-stranded cDNA made from poly-A^+ RNA by reverse transcription according to a method adapted from Kimmel and Berger (1987). Four lock-docking 1st strand cDNA primers were synthesized on an ABI model 381A DNA synthesizer (Applied Biosystems Inc., Mississauga, Ontario, Canada), in the format $T_{(12)}NX$, where a mixture of all four nucleotides was used for the penultimate position "N", and the

3' position "X" was one of either dA, dG, dT or dC. An equal mixture of these four oligonucleotides, at a final concentration of 500 μg/mL, was used to prime cDNA synthesis in a 50 μL reaction mixture containing 300 U of Superscript II reverse transcriptase (GibcoBRL), and 500 ng of mRNA. The initial reaction mixture also contained 1.0 μL of 2 mM non-labelled dATP, dGTP and dTTP, and 5.0 μL of 10 mCi/mL α-^{32}P labelled dCTP (Amersham). After 20 min at 42 °C, a chase of 1.0 μL of a solution containing all four non-labeled deoxyribonucleotides, each at a concentration of 10 mM, was added and the reaction was incubated for another 30 min at 42 °C. The resulting radiolabeled cDNA was immediately purified using a NucTrap probe purification column (Stratagene).

Hybond-N membranes carrying PCR-amplified cDNA were prehybridized at 60°C for 1 h in a buffer containing 7% SDS, 1 mM EDTA, 0.25 M NaH_2PO_4, and 1% BSA, and then hybridized with the labeled cDNA probe for 16 h at 60°C in the same buffer. Following hybridization, the membranes were washed in 2x SSPE (0.3 M NaCl, 0.02 M NaH_2PO_4, 2 mM EDTA) and 0.1% SDS for 10 min at room temperature, 1x SSPE and 0.1% SDS for 15 min at 65 °C, and finally in 0.1X SSPE and 0.1% SDS for 10 min at 65 °C. Autoradiography was performed using Biomax MR X-ray film (Kodak, Mississauga, Ontario, Canada) for 4 or 16 hours at -80°C, the autoradiograms were scanned using an Arcus scanner (Agfa, Toronto, Ontario, Canada), and the resulting images were analyzed for signal intensity using the public domain NIH Image program (developed at the US National Institutes of Health and available on the Internet at http://rsb.info.nih.gov/nih-image). The lowest detectable signal and the strongest signal were used to establish a range, and the range was divided into ten "bins" of equal size, providing expression level rankings from 1 to 10. For each clone, the 0.3 and 1.0 μg slot peak heights were averaged and this average value was used to assign the clone to the appropriate bin.

Northern Blotting and Hybridization

RNA was separated on denaturing 1% agarose gels containing 6% formaldehyde (Lehrach et al. 1977) and transferred to Hybond-N$^+$ (Amersham) membranes overnight by capillary blotting in 6x SSC. The hybridization probe was a gel-purified 235 bp PCR product representing the 3' portion of the coding region, prepared from 1st strand cDNA using the primer pair Tp177.1/Tp177.2, and labeled with the Prime It-RMT kit (Stratagene) in the presence of 5 μL of 10 mCi/mL α-^{32}P-labelled dCTP (Amersham). Probe purification, hybridization and autoradiography were carried out as described above for the RNA abundance analysis procedure.

Long Range-inverse PCR

Long range-inverse PCR was performed as described in Benkel and Fong (1996). Briefly, *T. pyriformis* genomic DNA was digested to completion with one of the following six-cutter restriction enzymes: *Bam*HI, *Eco*RI, *Hin*dIII, *Sac*I, *Spe*I, *Kpn*I or *Xba*I. The DNA fragments were circularized by ligation at a low DNA concentration (less than 1 μg/mL) and used as substrate for long range-inverse PCR (LR-IPCR) reactions with primer pairs listed in Table 2. The primary LR-IPCR reactions were followed up with secondary amplifications using pairs of nested PCR primers for the target site. For example, the primer pair Tp177.3/Tp177.4 was used for the primary LR-IPCR reaction for the isolation of genomic DNA fragments of the Tp177 gene. In this case the primary reaction was followed up with a secondary reaction in which a small aliquot of the Tp177.3/Tp177.4 reaction product was used as substrate in a PCR reaction with the 'nested' primer pair Tp177.5/Tp177.6. The resulting DNA fragments were subcloned into the vector pGEM-T for plasmid preparation and sequence analysis.

DNA Sequencing

Sequencing reactions were prepared using the BigDyeTerminator Cycle Sequencing Reaction Readimix (Applied Biosystems) and the extension products separated on an ABI 377 automated sequencer. Sequence analysis was carried out using Sequencher software (Gene Codes, Ann Arbor, Michigan, USA), and searches for similarity to known sequences in GenBank (http://www.ncbi.nlm.nih.gov) were carried out using Blast software (Altschul et al., 1990).

Promoter Analysis

The promoter comparison was limited to 422 bp since this represented the extent of the promoter sequence available for the *T. pyriformis* ubiquitin gene. Alignment of the promoter regions of a set of *T. pyriformis* and *Tetrahymena thermophila* genes was performed using ClustalW (Thompson et al., 1994) embedded within the BioEdit sequence analysis program (Hall, 1999). In a separate analysis, promot-

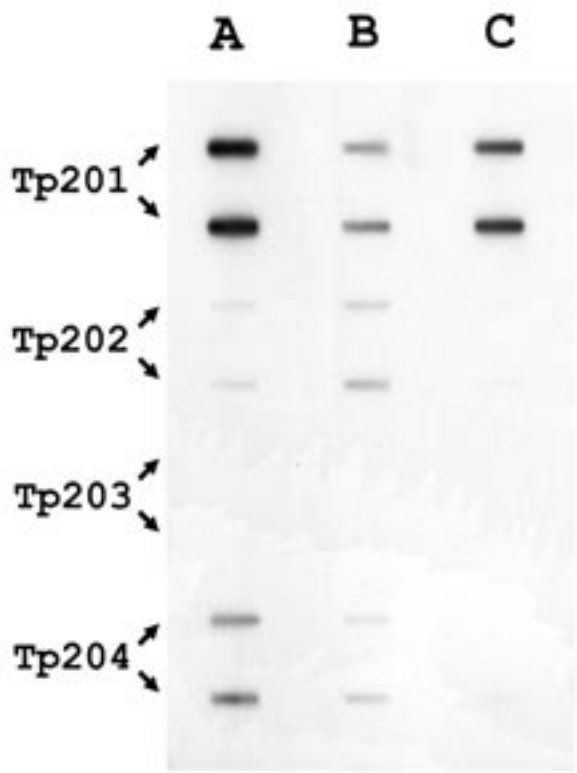

Fig. 1: Array of cloned *Tetrahymena* cDNAs hybridized with radiolabeled first-strand cDNA.
Clones in column 'A' are labelled; for each clone 0.3 and 1.0 μg of DNA was loaded in the upper and lower slots, respectively. A signal just above background, e.g. Tp202, was assigned a score of one (1) whereas clone Tp201 was ranked as an eight (8). Clones such as Tp203 that gave readings below the detection threshold of the assay were scored as zeros (0).

ers were scanned for transcription factor binding sites using two different software programs: (1) the Genomatix MatInspector program (www.genomatix.de); and (2) the EMBL-EBI Tfscan program accessible via the SRS/Lion portal at (www.ebi.ac.uk). In both cases, searches were conducted for transcription factor binding sites contained within the 'fungi' and 'other' components of databases, with no mismatches allowed.

Sequences used for the analysis were derived as follows: (1) *T. pyriformis* ubiquitin gene (X61053; Neves et al., 1991); (2) *T. thermophila* ubiquitin gene (U46561; Guerreiro and Rodrigues-Pousada, 1996); (3) *T. pyriformis* alpha tubulin gene (X57264; Soares et al., 1991); *T. thermophila* alpha tubulin gene (M86723; McGrath et al., 1994); *T. pyriformis* beta tubulin gene (X57265; Soares et al., 1991); *T. pyriformis* Tp177 gene (AY179365; this study); *T. thermophila* orthologue of Tp159 gene (CA916762; this study), hereafter called Tt159. The Tp159 EST represents a portion of the *T. pyriformis* orthologue of the *Arabidopsis thaliana* ribosomal L* (RPL8B) gene (see Table 1). A two-step process was used to identify the *T. thermophila* promoter sequence for this gene. Firstly, the *T. pyriformis* Tp159 sequence was aligned with the *A. thaliana* sequence to establish the location of the translation start codon. Next, the Tp159 sequence was compared to the TIGR (www.tigr.org) *T. thermophila* genome assembly (February, 2005) to identify the location of the coding region for this gene, and the 422 bp immediately upstream of the ATG uploaded into a BioEdit file to produce the Tt159 sequence. A similar process was used to extend the sequence information for the *T. thermophila* ubiquitin gene from the 252 bp included in U46561 to 422 bp.

Results and Discussion

CopyDNA Library Construction

Total RNA was extracted from *Tetrahymena pyriformis* cells grown in rich medium (SSP) under standard laboratory conditions, and poly-A+ RNA prepared for use as template for 1st strand cDNA synthesis. PCR amplified, double-stranded cDNA was ligated to the vector pGEM-T, and the resulting recombinant plasmids transformed into *E. coli* DH5α competent cells in order to construct a cDNA library for *T. pyriformis*.

RNA Abundance Analysis

In order to identify *Tetrahymena* genes that demonstrate robust expression levels, cDNA inserts were amplified from 153 clones by PCR, and the resulting products slot-blotted onto nylon filters and hybridized with radiolabeled first-strand cDNA. Two concentrations of PCR-amplified cDNA (0.3 μg or 1.0 μg per slot), both gauged to represent a large molar excess over any single component of the complex cDNA probe, were used for each clone to verify that the signal for any individual filter-bound sequence was proportional to the abundance of the transcript in the RNA pool. A representative slot blot showing a typical range of signal intensities is shown in Fig. 1. One hundred and seven clones showed expression above background in the array assay, and ten of these clones showed relatively high levels of expression (a relative rank of 5 or greater; see the "Expression" column in Table 1).

Table 1. *Tetrahymena pyriformis* EST clones

Clone	Protein Match[1]	E-value[2]	Species	Accession	Expression[3]
Tp1 / CA916755	ribosomal protein L23	4e-10	*D. melanogaster*	NP_523813	ND
Tp3 / CA916793	unknown				6
Tp4 / CA916743	unknown				ND
Tp6 / CA916744	isocitrate lyase	9e-21	*A. pinta*	Q43097	ND
Tp7 / CA916741	ribosomal protein L5	5e-33	*H. sapiens*	BAD92217	ND
Tp9 / CA916750	adenylate kinase B	6e-35	*P. tetraurelia*	CAH03446	ND
Tp112 / CA916747	unknown				4
Tp113 / CA916778	unknown				3
Tp114 / CA916796	granule lattice protein precursor (GRL5)	5e-27	*T. thermophila*	AF031321	3
Tp115 / CA916765	unknown				2
Tp116 / CA916766	unknown				3
Tp117 / CA916767	unknown				1
Tp118 / CA916768	putative ribosomal protein L39	5e-13	*A. thaliana*	NM_111086	6
Tp119 / CA916769	extracell. matrix receptor	1e-49	*P. carinii*	U09451	5
Tp120 / CA916770	unknown				0
Tp121 / CA916771	unknown				1
Tp122 / CA916772	unknown				1
Tp123 / CA916773	unknown				ND
Tp124 / CA916774	unknown				2
Tp125 / CA916775	unknown				ND
Tp126 / CA916776	unknown				3
Tp127 / CA916784	unknown				0
Tp128 / CA916783	unknown				0
Tp130 / CA916782	unknown				2
Tp131 / CA916781	unknown				1
Tp132 / CA916780	unknown				2
Tp133 / CA916779	unknown				2
Tp134 / CA916785	unknown				5
Tp135 / CA916787	ribosomal protein L29	3e-84	*T. thermophila*	M76718	5
Tp136 / CA916786	triose phosphate isomerase	3e-20	*S. cereale*	Z26875	3
Tp137 / CA916788	ribosomal protein S18	2e-29	*C. destructor*	AY014897	4
Tp138 / CA916789	unknown				2
Tp139 / CA916797	histone H3	4e-81	*T. thermophila*	M87504	3
Tp140 / CA916790	cellular differentiation-specific	1e-14	*T. vorax*	AF003090	2
Tp141 / CA916791	unknown				2
Tp142 / CA916792	fructose 1,6-bisphosphate aldolase	1e-37	*C. caldarium*	AF217804	3
Tp155 / CA916759	ribosomal protein S10	9e-11	*S. pombe*	NP_595605	6
Tp159 / CA916760	ribosomal protein L8	5e-56	*A. thaliana*	NM_114978	4
Tp163 / CA916757	ribosomal protein S15	5e-36	*R. raetam*	AF439281	4
Tp167 / CA916777	overlaps Tp177				8
Tp168 / CA916764	Csf-1	6e-19	*C. sativus*	AB008846	4
Tp177 / CA916762	unknown				6

Tp199 / CA916761	unknown				5
Tp201 / CA916763	unknown				10
Tp317 / CA916756	triose phosphate isomerase	8e-26	*H. vulgare*	U83414	ND
Tp318 / CA916794	unknown				ND
Tp319 / CA916758	unknown				ND
Tp320 / CA916754	ribosomal protein L33	3e-26	*C. sativa*	AF334840	ND
Tp321 / CA916753	translation elongation factor 2	8e-21	*T. thermophila*	AAN04122	ND
Tp336 / CA916752	ribosomal protein L17	1e-10	*C. paradoxa*	CAB56830	ND
Tp 338 / CA916749	ribosomal protein L18	1e-24	*D. discoideum*	XP_641456	ND
Tp339 / CA916748	dynein light chain 4	1e-10	*S. japonicum*	AF072330	ND
Tp341 / CA916745	unknown				ND
Tp342 / CA916746	importin alpha	3e-38	*P. tetraurelia*	CAH03230	ND
Tp343 / CA916795	unknown				ND
Tp 345 / CA916751	elongation factor 2	1e-113	*T. pyriformis*	AF213665	ND
Tp 347 / CA916742	ribosomal protein L5	3e-32	*H. sapiens*	AAP06189	ND

[1] Sequences were compared to GenBank using the tBlastx program (Altschul et al., 1990) to determine sequence similarity with known genes.

[2] Only matches with an E-value of less than 1e-09 are shown.

[3] The signal intensities from the slot-blot hybridization experiments were ranked on an arbitrary scale from 0 to 10, where a score of "10" was assigned to the highest level of signal detected and "0" indicates that the signal strength was below the detection threshold of the assay. An additional 96 clones were analyzed for mRNA abundance, but sequence data have for these genes have not yet been obtained. "ND" indicates clones for which single pass sequence was derived, but which were not included in the expression analysis.

Expressed Sequence Tags

Fifty-seven cDNA clones, including all ten high level expressors identified by array hybridization, were subjected to single pass DNA sequence analysis. Following the removal of vector sequences and trimming to a minimum quality value of Phred 20 (Ewing and Green, 1998), the cDNA sequences ranged from 131 to 638 bp in length, with an average length of 412 bp, and displayed an average A+T% of 65 (52% to 84%). These sequences have been submitted to GenBank (accessions CA916741 to CA916797) as a set of *T. pyriformis* expressed sequence tags (EST); their relationships to known sequences are shown in Table 1. Eleven ESTs, or 19%, were similar to ribosomal protein genes reported for other species. Support was relatively strong for some of the assignments, e.g. between clone Tp135 and the *T. thermophila* L29 gene with an E-value of 3e-84, and weaker for others, e.g. between clone Tp1 and the *D. melanogaster* L23 gene with an E-value of 4e-10. The predicted translation products of fifteen EST sequences matched known proteins in GenBank, including a cellular differentiation factor (Tp140), an isocitrate lyase (Tp6), and histone 3 (Tp139). The remaining thirty-one ESTs showed no compelling similarity to known protein sequences.

Two clones, Tp177 and Tp201, with expression levels of six and eight respectively, were chosen for further study. The complete sequences of these two cDNA inserts were derived and compared to the non-redundant GenBank database using the tBlastx algorithm. The results showed no significant matches for either sequence.

Characterization of Highly Expressed Genes

Cloning of the genomic sequences flanking the transcribed regions for clones Tp177 and Tp201 was carried out using the long range-inverse PCR technique (LR-IPCR; Benkel and Fong, 1996) with primer pairs designed using cDNA sequences. Attempts to obtain amplified LR-IPCR products for clone Tp201 were unsuccessful. In contrast, a contig spanning the entire Tp177 gene was assembled following three rounds of LR-IPCR.

The first round of LR-IPCR on DNA circles from

*Eco*RI-digested genomic DNA using Tp177 cDNA sequence-based primer pairs Tp177.3/Tp177.4 followed by Tp177.5/Tp177.6 (for primer sequences see Table 2) yielded a fragment of approximately 950 bp representing the downstream portion of the gene. A second round of LR-IPCR with primer pairs Tp177.7/Tp177.8 followed by Tp177.9/Tp177.10 produced an *Eco*RI fragment of 650 bp, which revealed the presence of a *Spe*I restriction site in the upstream portion of the cDNA sequence. A third round of LR-IPCR performed using *Spe*I-digested DNA and primer pairs Tp177.11/Tp177.12 followed by Tp177.13/Tp177.14 yielded a product of approximately 1,650 bp, which extended the existing sequence by roughly 1,500 bp in the upstream direction. Assembly of the sequences derived through analysis of the cDNA and LR-IPCR clones for this gene resulted in a contig of 3,064 bp that spanned the entire functional gene including upstream, protein encoding, and downstream sequences (Fig. 2). The accuracy of the configuration deduced from the LR-IPCR-derived fragments was tested by amplifying the bulk of the Tp177 gene directly from *T. pyriformis* genomic DNA using a pair of PCR primers designed according to the sequence of the LR-IPCR-derived contig. The DNA sequence of the genomic DNA-derived amplicon confirmed the accuracy of the configuration shown in Figure 2.

Since Tp177 is unlike any other previously characterized gene, it was not possible to use an orthologue as a guide in determining gene structure. Instead, the size of the Tp177 transcript was determined and the putative transcription start site mapped, as described below. Firstly, the size of the Tp177 transcript was determined by Northern analysis. Samples of *T. pyriformis* total and poly-A$^+$ RNA were separated on a denaturing gel and the resulting blot hybridized with radiolabeled Tp177 cDNA (see Materials and Methods). Autoradiography revealed a Tp177-specific transcript of approximately 1.65-1.75 kb (Fig. 3). Secondly, the transcription start site was localized by performing PCR on 1st strand cDNA in a series of amplifications where a common lower strand (downstream) primer Tp177.2, was used in combination with one of five upper strand primers, Tp177.22, Tp177.23, Tp177.26, Tp177.24 and Tp177.25, that were designed to anneal at distances of 1,013, 1,420, 1,553, 1,693 and 1,863 n upstream of the Tp177 poly-adenylation site, respectively (see Fig. 2). PCR reactions were conducted in parallel on both genomic DNA and 1st strand cDNA. A PCR product was obtained in every case when genomic DNA was used as template. In contrast, PCR products were obtained from the cDNA template for the primers Tp177.2, Tp177.3 and Tp177.26 only – not when primers Tp177.24 or Tp177.25 were used in combination with Tp177.2. Moreover, for those primer pairs that yielded PCR products on both genomic DNA and 1st strand cDNA, the amplified fragments for both substrates were identical in size within the resolution of the agarose gel assay. Thus, if the Tp177 gene contains introns they are either very small (less than 100 bp in total) or located in the extreme 5'-nontranslated portion of the transcribed region. Both the Northern and cDNA analysis mapped the transcriptional start site of the Tp177 gene to the region of roughly 100 bp between the annealing sites of oligomers Tp177.24 and Tp177.26 (Fig. 2).

Table 2. Sequences of oligonucleotides used for amplification of cDNAs and for long range-inverse PCR and transcription start site mapping on *Tetrahymena* clone Tp177

PRIMER	SEQUENCE (5' to 3')
SMART	AAGCAGTGGTAACAACGCAGAGT
Tp177.1	CATACTCCCAAGTTGCCGATC
Tp177.2	TAAGAAGGGAATGTGCGAACAC
Tp177.3	CCTAGAGTAGACGAGCTACTGG
Tp177.4	AAGATATAGGTCATGTGTTCGC
Tp177.5	GAGCTACTGGTCGATTATACTC
Tp177.6	TAGGTCATGTGTTCGCACATTC
Tp177.7	TATTATTAGATCGGCAACTTGG
Tp177.8	GGTCATCAATCAATCTAAATTG
Tp177.9	CGGCAACTTGGGAGTATGTTGG
Tp177.10	TCTAAATTGCCCTCTGCGAACC
Tp177.11	GCTGGCGTGTTTCTGTAATAGG
Tp177.12	CCAACAATCCGTTCACCGAATG
Tp177.13	CAATACTTAGTGAATTGCTGTG
Tp177.14	CTCTCAATGGAAATACTGATGC
Tp177.22	CATACACTCCGTTTCTAACTTC
Tp177.23	CTTGCTTTTGAGACTAAATAACC
Tp177.24	CACCACCCAAACGAATATAATC
Tp177.25	GTTCAATTATCAGGCAGGCTG
Tp177.26	CGGAAAAGAAAGGTGAGTATTTC

Analysis of the Tp177 transcript revealed a number of open reading frames (ORFs), including several of 50 amino acids or less. The longest ORF of 378 bp encodes a protein of 126 amino acid residues which yielded no significant matches to known proteins or conserved domains when compared to the relevant NCBI databases with the RPS-BLAST and reiterative PSI-BLAST programs (Altschul et al., 1997).

Overall, the Tp177 gene shows an A+T bias of approximately 68%. As far as specific portions of the

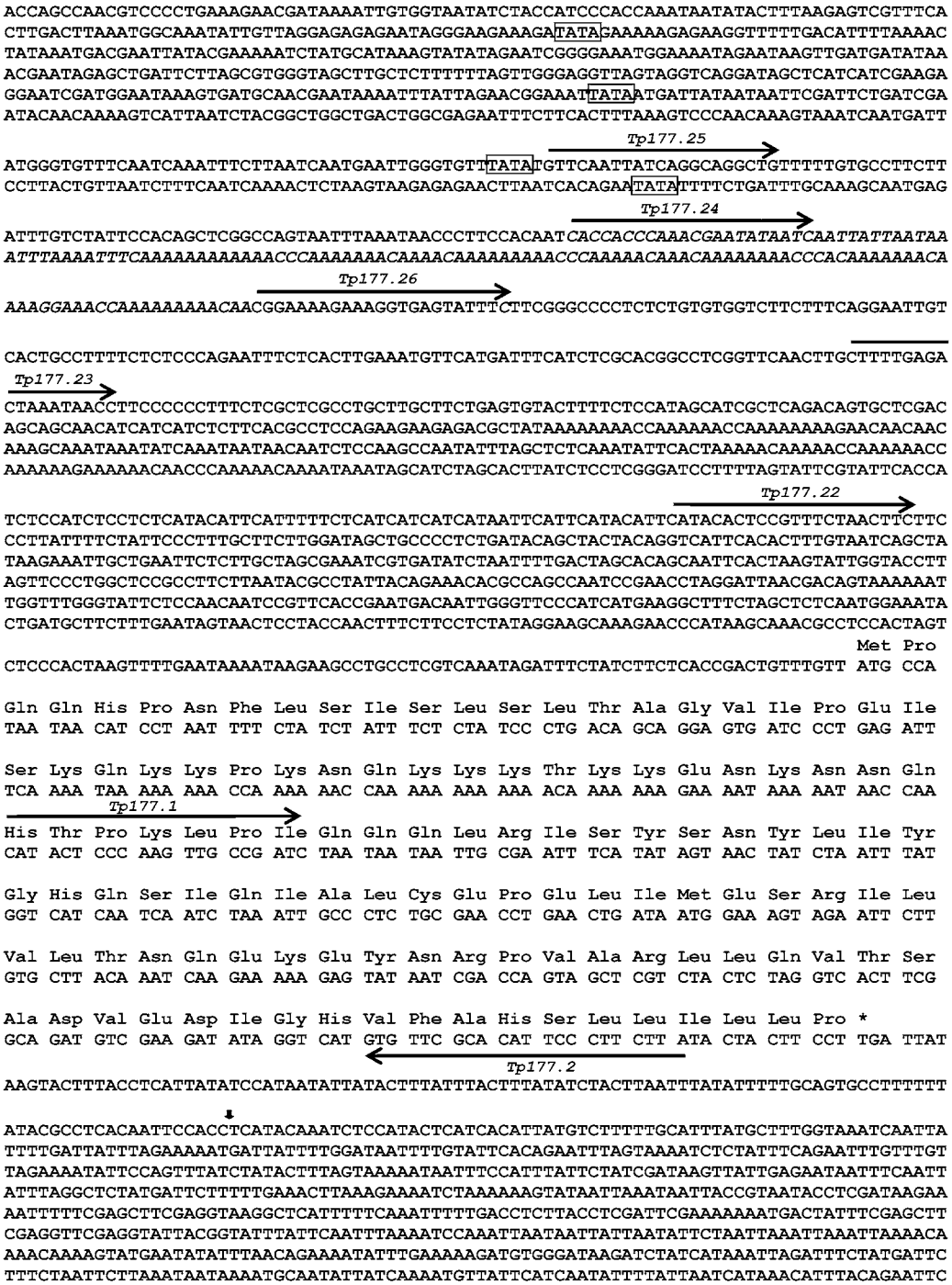

Fig. 2: Nucleotide sequence of the Tp177 gene. The amino acid sequence of the predicted protein encoded by the longest open reading frame in the Tp177 transcript is shown above the coding sequence. The annealing sites for oligonucleotides used to localize the transcription start site are shown as horizontal arrows either above (upper strand primers) or below (lower strand primer) the nucleotide sequence. The transcription start site was mapped to a 100 bp region (shown in italics) flanked by the annealing sites for the primers Tp177.24 and Tp177.26. TATA-box motifs in the putative promoter region are boxed, and the polyadenylation site is indicated with a vertical arrow. The Tp177 sequence is available from GenBank as accession number AY179365.

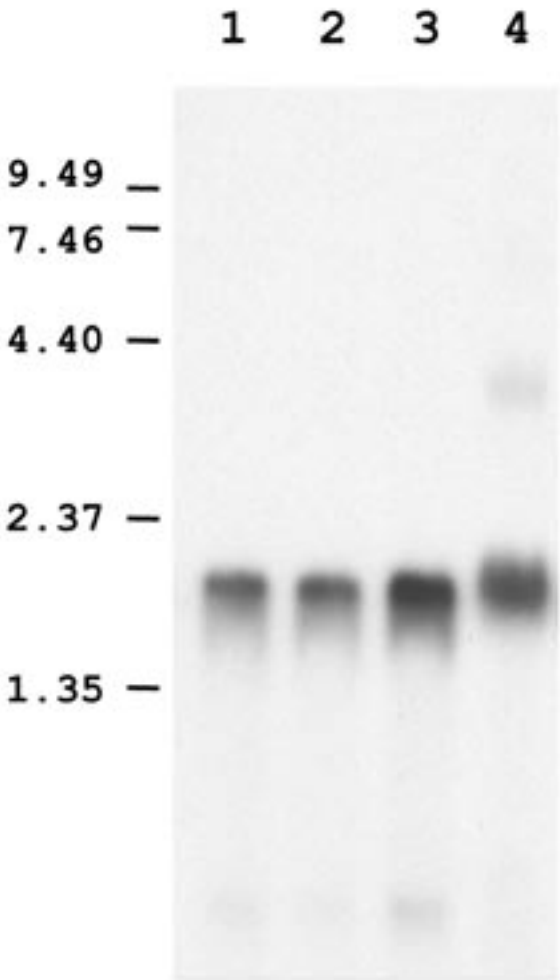

Fig. 3: Northern analysis of the Tp177 transcript. Northern blot containing approximately 5 ug of poly-A+ RNA (lanes 1 & 2), 10 ug of poly-A+ RNA (lane 3) and 30 ug of total RNA (lane 4) hybridized with a Tp177-specific probe. The size of the most abundant Tp177-related transcript was estimated at approximately 1.7 kb. Positions of size standards are marked (in kb).

gene are concerned, the putative 5' non-coding region of the Tp177 transcript, the coding region, and sequence downstream of the stop codon are 63%, 66%, and 79% A+T, respectively, whereas the putative gene promoter of roughly 700 bp (Fig. 2) is 67% A+T, including an C+A-rich stretch surrounding the putative transcription start site.

Promoter Analysis

The results described above on gene-specific RNA expression levels indicate that the Tp177 gene is robustly expressed in *T. pyriformis* cells. In addition, as a first step in the functional analysis of the Tp177 promoter, upstream sequences of the Tp177 gene have recently been used to drive the expression of a luciferase reporter gene in transformed *Saccharomyces cerevisiae* cells. (Benkel et al., 2007). Taken together, these results suggest that a DNA fragment representing the Tp177 promoter region could be used directly in hybrid gene constructs to drive the robust expression of foreign genes in transformed *Tetrahymena* clones. An alternative approach for the high level production of recombinant peptides would involve the further enhancement of useful 'wild type' *Tetrahymena* promoters (Shang et al., 2002) or the construction of fully-synthetic 'designer' promoters optimized for the production of specific end products, using DNA motifs for functional regulatory elements. In order to gain insight into the composition of *Tetrahymena* promoters, we conducted an analysis using a set of seven promoter sequences currently available from public databases, which allowed for comparisons within *T. pyriformis* as well as intergenus comparisons between *T. pyriformis* and *T. thermophila* (see Materials and Methods for a full description).

Overall, alignment of the promoter regions of a set of seven *Tetrahymena* genes revealed moderate sequence identity levels ranging from a low of 37% to a high of 58%. The maximal level of conservation was observed between the *T. pyriformis* and *T. thermophila* alpha tubulin promoters (58% identity) followed by the *T. pyriformis* alpha and beta tubulin promoters (49%). The lowest level of sequence similarity (37%) was observed between two strong promoters identified in this study, i.e. the *T. pyriformis* Tp177 gene promoter and the *T. thermophila* equivalent of the Tp159 gene promoter, Tt159 (see Materials and Methods). The sequence identity of the *T. pyriformis* and *T. thermophila* alpha tubulin genes was highest within a block of 145 bp immediately upstream of the ATGs (73%). Sequence identity between the *T. pyriformis* alpha and beta tubulin promoters over this same stretch was 61%. Sequence similarity dropped markedly upstream of the 145 bp mark in both comparisons; for example, dropping from 73% for the first 145 bp to 52% for the remainder of the 422 bp included in the comparison for the inter-species alpha tubulin promoter comparison. Alignment of the *T. pyriformis* and *T. thermophila* ubiquitin promoters failed to reveal any sizeable regions of elevated sequence conservation.

Next, a more detailed analysis was performed using two popular analysis programs, Tfscan (Wingender, 1988) and MatInspector (Cartharius et al., 2005), which are designed to identify transcription factor binding sites within promoter sequences.

All seven promoter sequences contained at least one putative TATA-box. The Tp177 promoter contained three TATA-boxes, whereas the Tt159 promoter contained five TATA motifs, the highest number of any of the promoters included in the analysis. Both Ftscan and MatInspector identified multiple potential transcription factor binding motifs in every promoter tested. Common motifs included recognition sites for the NIT2 activator of nitrogen-regulated genes (TATC; present in 6 / 7 genes), the GAL4 transcription activator (GATAA; 6 / 7 genes), and the GCN4 activator (TGAGTG; 3 / 7 genes). Although both the Tp159 and Tp177 ESTs show robust levels of expression as measured by macroarray analysis, the proximal promoter sequences of Tp177 and Tt159 (the *T. thermophila* equivalent of Tp159) are quite different as far as putative transcription factor binding site content is concerned. The Tt159 promoter contained several copies of the GAL4 (GATAA) motif. In contrast, the Tp177 promoter contained RAP1, ADR1, and MADS-box motifs. Both promoters contained putative NIT2 and HIS4 binding sites. Although both promoters contained A-rich sequence spanning the putative transcription start sites, neither promoter contained the regulatory motifs previously identified in *T. thermophila* histone genes (Brunk and Sadler, 1990; Larsen and Kristiansen, 1995). Likewise, both promoters also lacked the basal level control element and the UV repressor/activator element of the *T. thermophila* RAD51 gene (Smith et al., 2004). However, the upstream region of the Tp177 gene contained two copies of RAD51 UV inducer motif (TTTCAAT) spaced roughly 100 bp apart.

The above analysis must be, however, interpreted with caution. Firstly, because of the lack of information on protist promoters in general, the transcription factor databases used for the promoter scan contained primarily yeast factor binding sites. Secondly, the A+T content of *Tetrahymena* genes is high on average; for example, the EST sequences derived in this study showed an average A+T content of 65%. This skewed base content may lead to spurious binding site predictions for transcription factors with A+T-rich recognition sites. As a result, the ability of comparative studies to dissect protist promoters is limited and reliable information on the regulatory components of *Tetrahymena* promoters will only be developed through functional evaluation of promoter sequences of the kind described in Smith et al. (2004).

Based on recent evidence indicating that *Tetrahymena* cells can produce authentic, correctly processed and glycosylated, human DNase I enzyme, Weide et al. (2006) suggested that *Tetrahymena* has considerable potential as a high-quality expression system for recombinant mammalian proteins. At the same time, they pointed to low yield of expression as one of two main areas that need to be addressed in order for the ciliate expression system to become competitive with alternative systems for the production of mammalian proteins. A number of factors can contribute to low yield of recombinant product including promoter strength, transcript turnover rate, and recombinant peptide stability. The goal of this study was to address the former issue by identifying gene promoters that direct high level gene expression in *Tetrahymena*. The Tt159 and Tp177 promoters identified in this study will aid in the analysis of promoter elements and the design and assembly of DNA constructs that promote high level production of foreign peptides, thereby allowing the full potential of *Tetrahymena* as a recombinant protein production system to be realized.

Acknowledgements

This work was supported by grants from Dairy Farmers of Ontario and Agriculture & Agri-Food Canada's Matching Investment Initiative. The authors thank B. Genswein, Lethbridge Research Centre of Agriculture and Agri-Food Canada, for conducting database searches with *T. pyriformis* ESTs.

References

Altschul S.F., Gish W., Miller W., Myers E.W. and Lipman, D.J. 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403-410.

Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W. and Lipman D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.

Benkel B.F. and Fong Y. 1996. Long range-inverse PCR (LR-IPCR): extending the useful range of inverse PCR. Genet. Anal. 13, 123-127.

Benkel B.F., Richmond S., Gusse J., Zhao Y., Ivan M., Forster R.J. and Teather R.M. 2007. Robust expression in yeast cells of a reporter gene driven by rumen protoroal promoter sequences. World J. Microbiol. Biotechnol. Publ. online.

Brunk C.F. and Sadler L.A. 1990. Characterization of the promoter region of *Tetrahymena* genes. Nucleic Acids Res. 18, 323-329.

Cartharius K., Frech K., Grote K., Klocke B., Haltmeier M., Klingenhoff A., Frisch M., Bayerlein M. and Werner T. 2005. MatInspector and beyond: promoter analysis based on transcription factor binding sites. Bioinformatics. 21, 2933-2942.

Cassidy-Hanley D., Bowen J., Lee J.H., Cole E., VerPlank L.A., Gaertig J., Gorovsky M.A. and Bruns P.J. 1997. Germline and somatic transformation of mating *Tetrahymena thermophila* by particle bombardment. Genetics. 146, 135-147.

Ewing B. and Green P. 1998. Base-calling of automated sequencer traces using Phred. II. Error probabilities. Genome Res. 8, 186-94.

Gaertig J. and Gorovsky M.A. 1992. Efficient mass transformation of *Tetrahymena thermophila* by electroporation of conjugants. Proc. Natl. Acad. Sci. USA 89, 9196-9200.

Guerreiro P. and Rodrigues-Pousada C. 1996. Characterization of a polyubiquitin gene in *T. thermophila* and of ubiquitin gene expression during reproduction and under stress conditions. Gene. 182, 183-188.

Hall T.A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. 41, 95-98.

Kimmel A.R. and Berger S.L. 1987. Preparation of cDNA and the generation of cDNA libraries: overview. Methods Enzymol. 152, 307-316.

Larsen L.K. and Kristiansen K. 1995. Transcription *in vitro* of *Tetrahymena* class II and class III genes. J. Biol. Chem. 270, 7601-7608.

Lehrach H., Diamond D., Wozney J.M. and Boedtker H. 1977. RNA molecular weight determinations by gel electrophoresis under denaturing conditions: A critical re-examination. Biochemistry. 16, 4743-4751.

McGrath K.E., Yu S.M., Heruth D.P., Kelly A.A. and Gorovsky M.A. 1994. Regulation and evolution of the single alpha-tubulin gene of the ciliate *Tetrahymena thermophila*. Cell Motil. Cytoskeleton. 27, 272-283.

Neves A.M., Guerreiro P., Miquerol L. and Rodrigues-Pousada C. 1991. Molecular cloning and expression of a *Tetrahymena pyriformis* ubiquitin fusion gene coding for a 53-amino-acid extension. Mol. Gen. Genet. 230, 186-192.

Orias E., Hamilton E.P. and Orias J.D. 2000. *Tetrahymena* as a laboratory organism: useful strains, cell culture, and cell line maintenance. In: Methods in cell biology, vol. 62, Eds. Asai, D.J. and Forney, J.D., Academic Press, New York, NY, pp 190-208.

Shang Y., Song X., Bowen J., Corstanje R., Gao Y., Gaertig J. and Gorovsky M.A. 2002. A robust inducible-repressible promoter greatly facilitates gene knockouts, conditional expression, and overexpression of homologous and heterologous genes in *Tetrahymena thermophila*. Proc. Natl. Acad. Sci. USA 99, 3734-3739.

Smith J.J., Cole E.S. and Romero D.P. 2004. Transcriptional control of *RAD51* expression in the ciliate *Tetrahymena thermophila*. Nucleic Acids Res. 32, 4313-4321.

Soares H., Cyrne L., Barahona I. and Rodrigues-Pousada C. 1991. Different patterns of expression of beta-tubulin genes in *Tetrahymena pyriformis* during reciliation. Eur. J. Biochem. 197, 291-292.

Takemasa T., Ohnishi K., Kobayashi T., Takagi T., Konishi K. and Watanabe Y. 1989. Cloning and sequencing of the gene for *Tetrahymena* calcium-binding 25-Da protein (TCBP-25). J. Biol. Chem. 264, 19293-19301.

Thompson J.D., Higgins D.G. and Gibson T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673-4680.

Weide T., Herrmann L., Bockau U., Niebur N., Aldag I., Laroy W., Contreras R., Tiedtke A. and Hartmann M.W. 2006. Secretion of functional human enzymes by *Tetrahymena thermophila*. BMC Biotechnol. 16, (in press).

Wingender E. 1988. Compilation of transcription regulating proteins. Nucleic Acids Res. 16: 1879-1902.

Address for correspondence: Bernhard F. Benkel. Department of Plant and Animal Sciences, Nova Scotia Agricultural College, P.O. Box 550, Truro, Nova Scotia, B2N 5E3, Canada. Tel.: 902-893-6165; FAX: 902-895-6734. E-mail: bbenkel@nsac.ca.

Editorial responsibility: Alastair Simpson